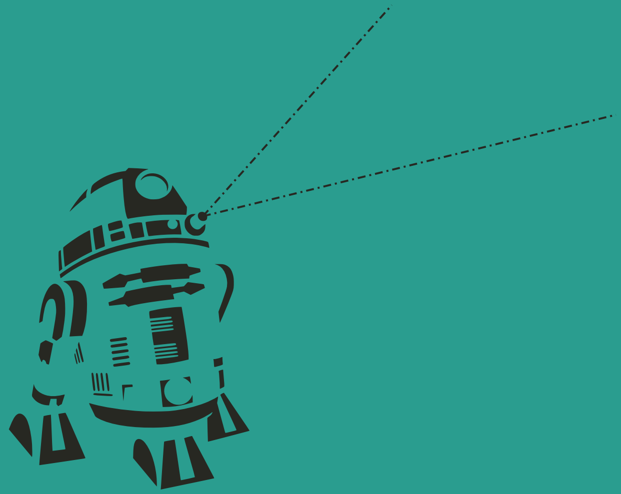
Introduction to Machine Learning

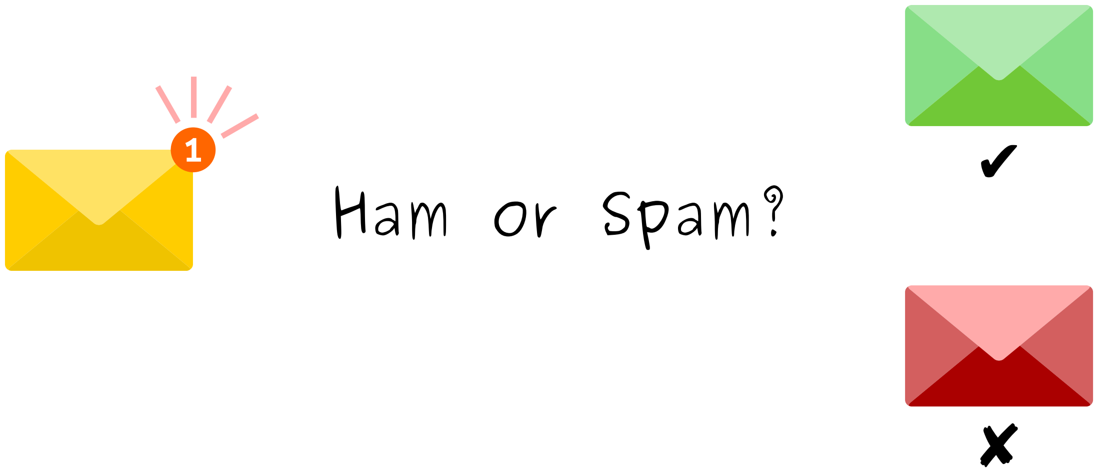
DSAA

Albert Ruiz

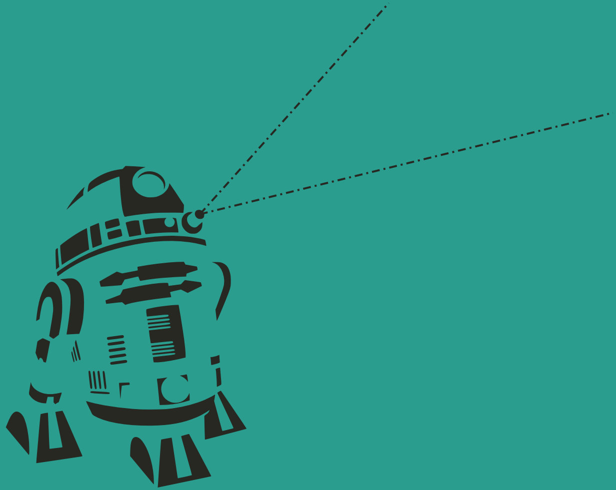
What was the first Machine Learning application?



First ML application: the spam filter



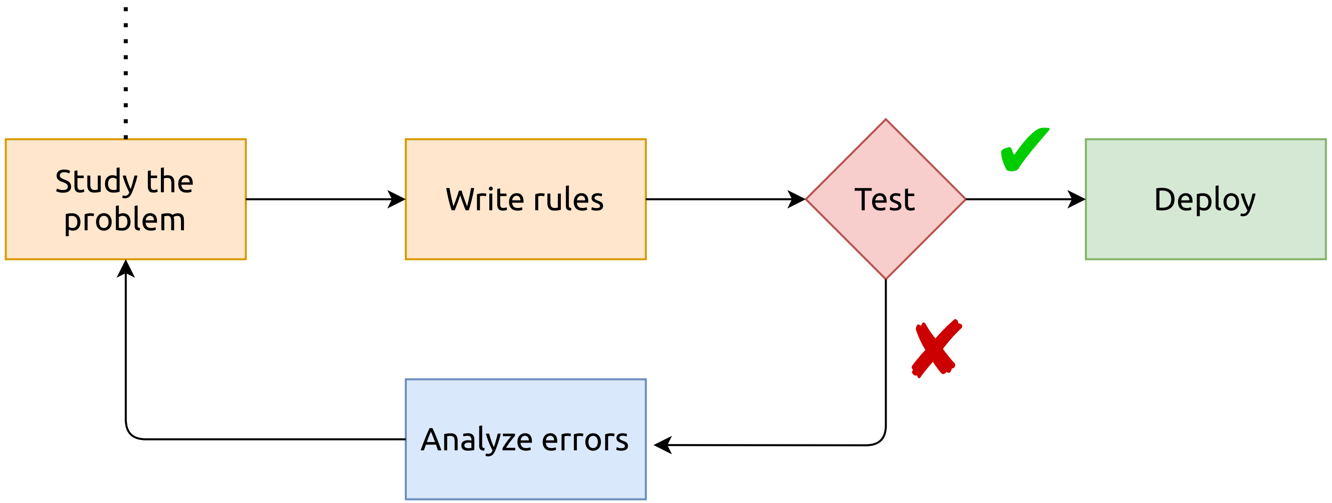
How would you code a spam filter?



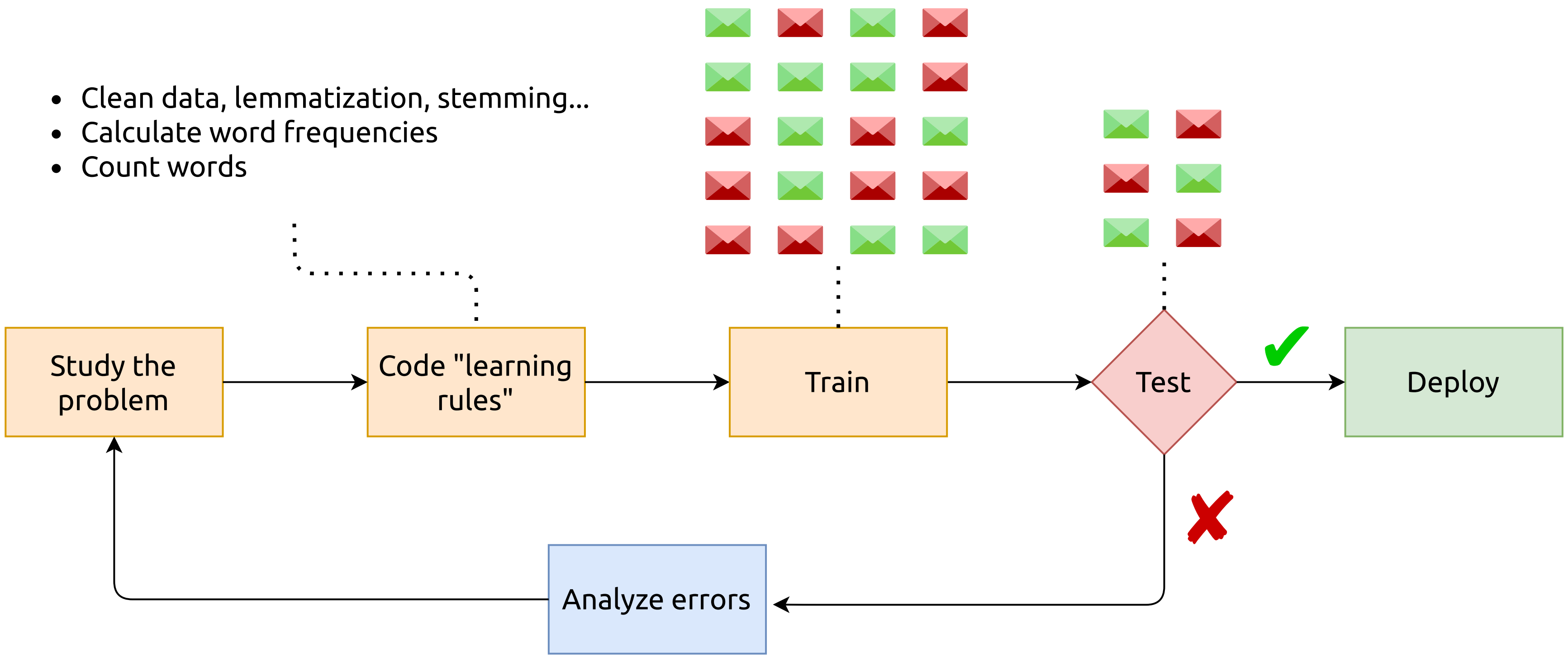
Traditional approach: the developer learns

Developer does (example):

- Find common words: IBAN , discount , offer , bank , password ...
- Find patterns in introductory phrases: Dear Sir/Madam , Mr/Mrs/Miss ...
- Find patterns in email addresses: @hacker.com , @no-reply.com ...
- Calculate weights



ML approach: the machine learns



What are we going to learn today?

Agenda

1. Introduction (5 min)
    - What is ML?
    - Why ML?
2. End-to-end ML (45 min)
    - Data
    - Processing
    - Feature extraction
    - Modelling
    - Results
3. Hands-on ML (practice, 1h)

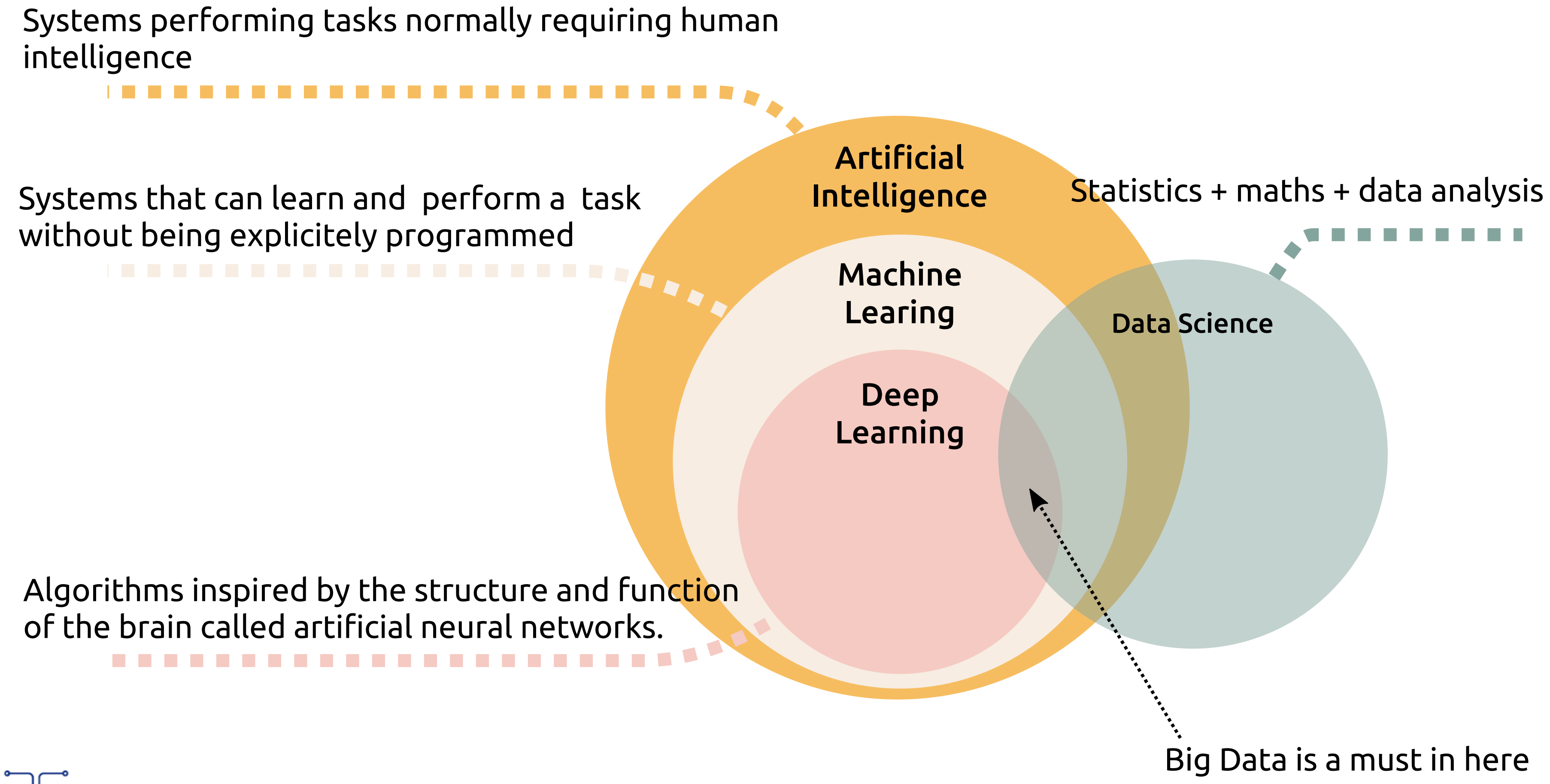
Introduction

What is ML?

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

[Arthur L. Samuel, 1959]

AI / ML / DL / DS / BD



ZURICH
Technology Delivery Center
ServiZurich I Barcelona

ML can help humans learn!

Some modern problems are too complex for traditional approaches:

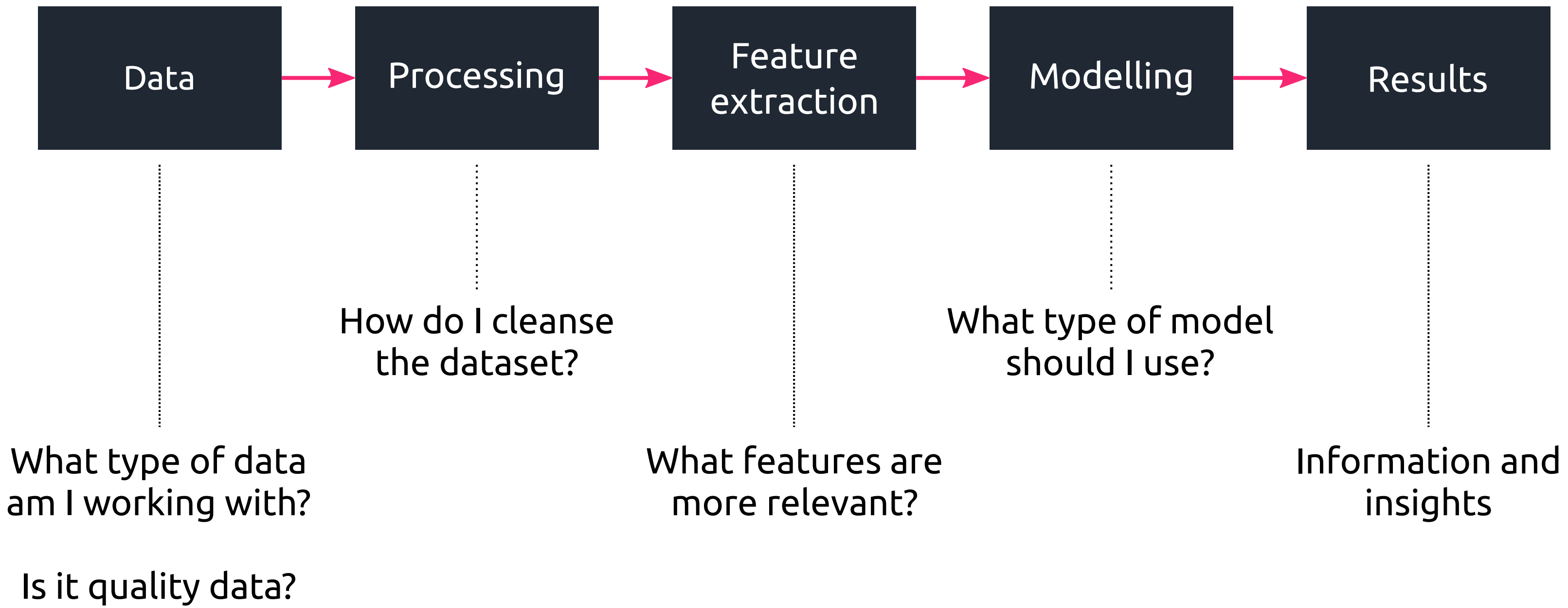
- Problems that require fine-tuning or long list of rules
- Problems with fluctuating data
- Getting insights from large amounts of data

A wide range of use cases

- Text classification
- Sentiment analysis
- Summarizing long text
- Data extraction from images
- Fraud detection
- Chatbots
- Client segmentation
- Recommending a product to a client
- Speech recognition
- Forecasting

Common steps in a ML project

The common steps



Data

Data

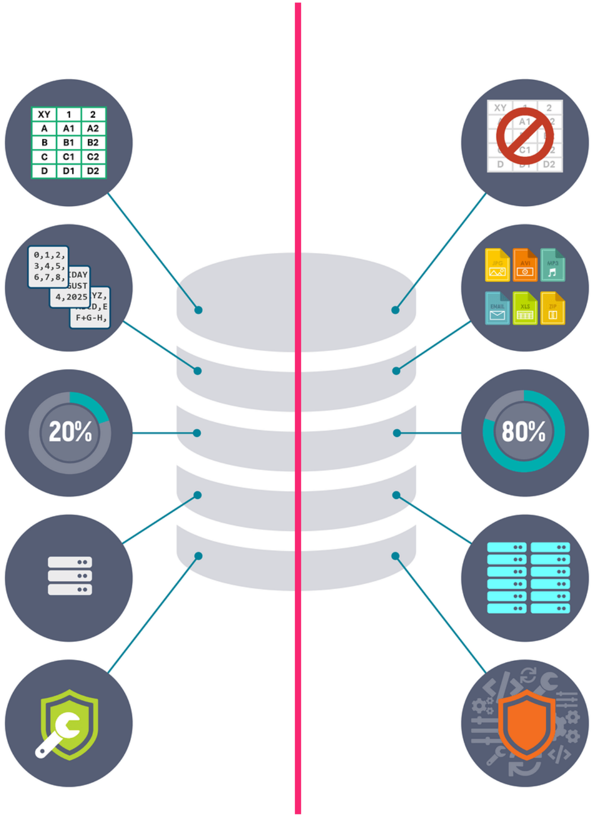
Processing

Feature extraction

Modelling

Results

Structured / Unstructured



Structured	Unstructured
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, emails spreadsheets...
Estimated 20% of enterprise data (Gartner)	Estimated 80% of enterprise data (Gartner)
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions

Labelled / Unlabelled

Labelled

Content	Email	Label
Dear Sir/Madam, we need to validate your user Id and password of your bank account...	me@noreply.com	spam
Hello, I am a rich businessman and I need help. My wallet was stolen in the airport...	florentino@rmad.com	spam
Hello, I can't find the ML presentation on teams. Can you please send me a link? BR	albert@zurich.com	ham
Congratulations! Because you are our client 1M, you have won the new Nintendo....	mike@gifts-tonight.com	spam
Hey buddy, I think today we have a meeting, right? I haven't received any invitation...	kim@megacom.com	ham
Your job alert for full stack engineer. 1 new job in LA matches your preferences.	jobalerts@linkedin.com	ham

Unlabelled

Data

Processing

Feature extraction

Modelling

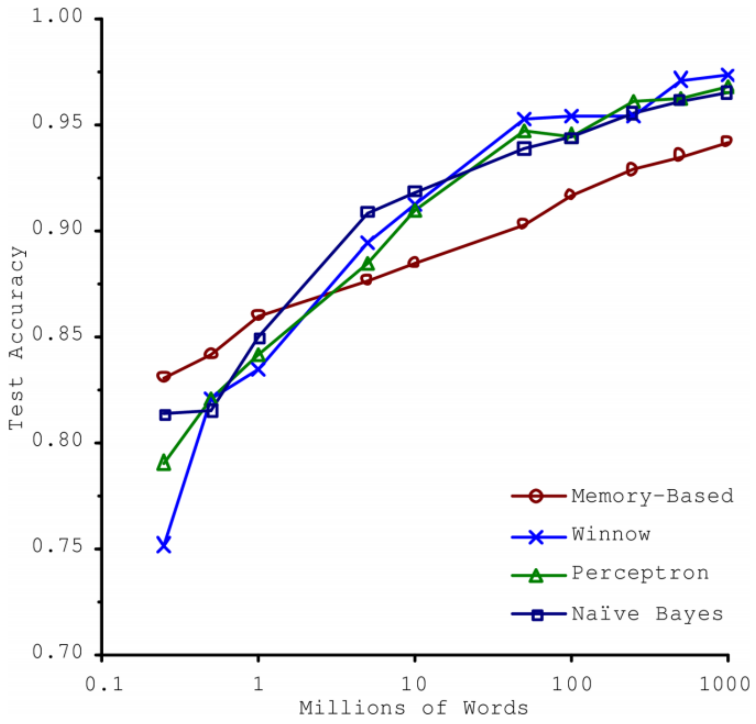
Results

Categorical / Quantitative

Categorical	Quantitative
Sex M, F	**Age** 18, 19, 20, 21, 22, 23 ...
Country England, France, Spain, Switzerland ...	**Temperature (ºC)** 18, 20.5, 22, 23,5, 25
Status OK, KO, UNKNOWN	**Income (k€)** 40.5, 55.0, 65.5, 150.0 ...
Brands Peugeot, Ferrari, SEAT, Audi, Hispano-Suiza ...	**Number of employees** 10, 15, 70, 323, 998 ...

ZURICH Technology Delivery Center ServiZurich I Barcelona

Invest more on data



Perform better when trained with more data

4 different algorithms

[Microsot article (2001)]

Data

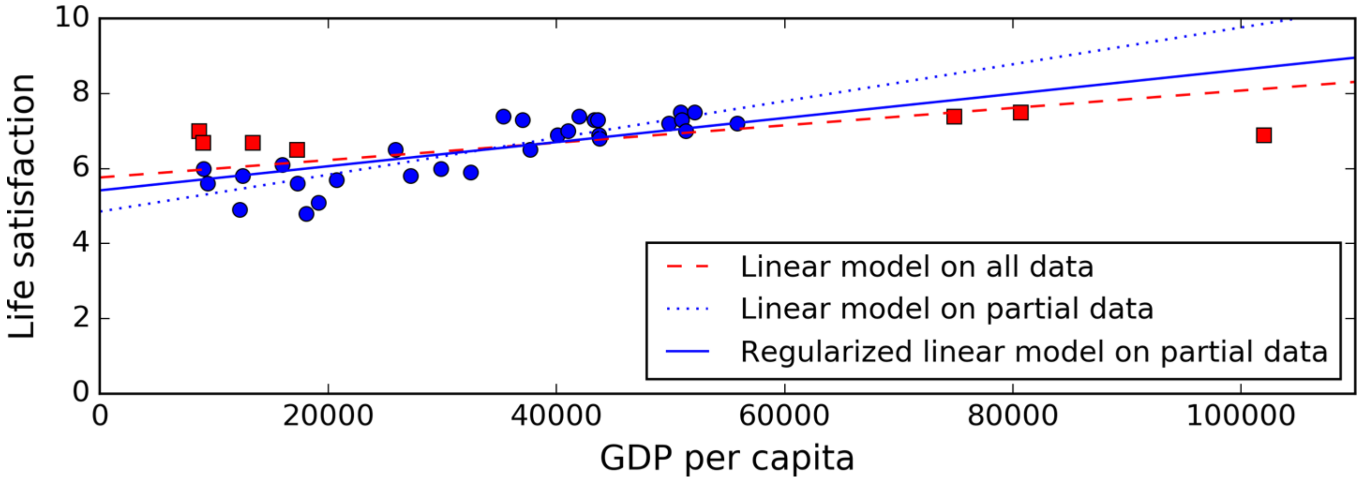
Processing

Feature extraction

Modelling

Results

And use all data!



Data

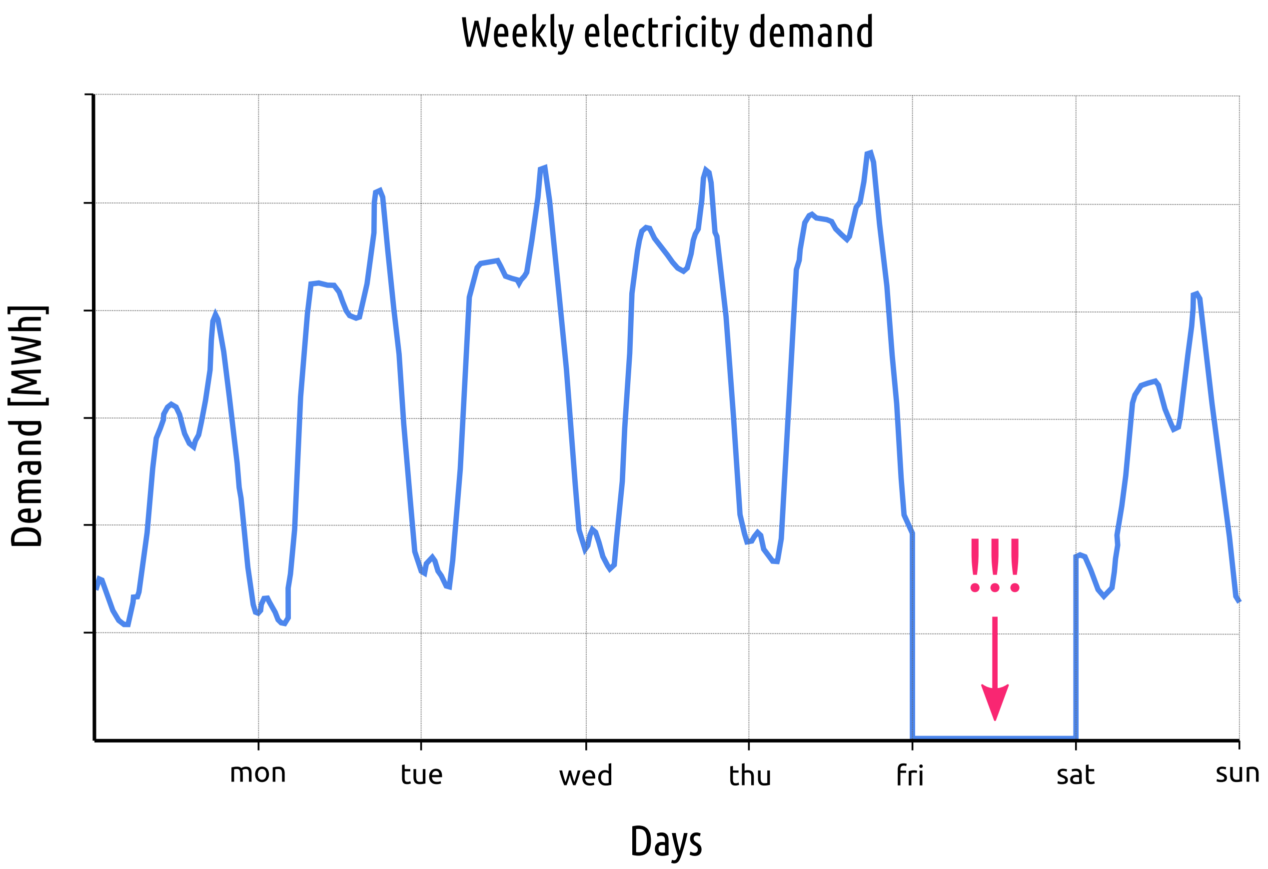
Processing

Feature extraction

Modelling

Results

Poor quality

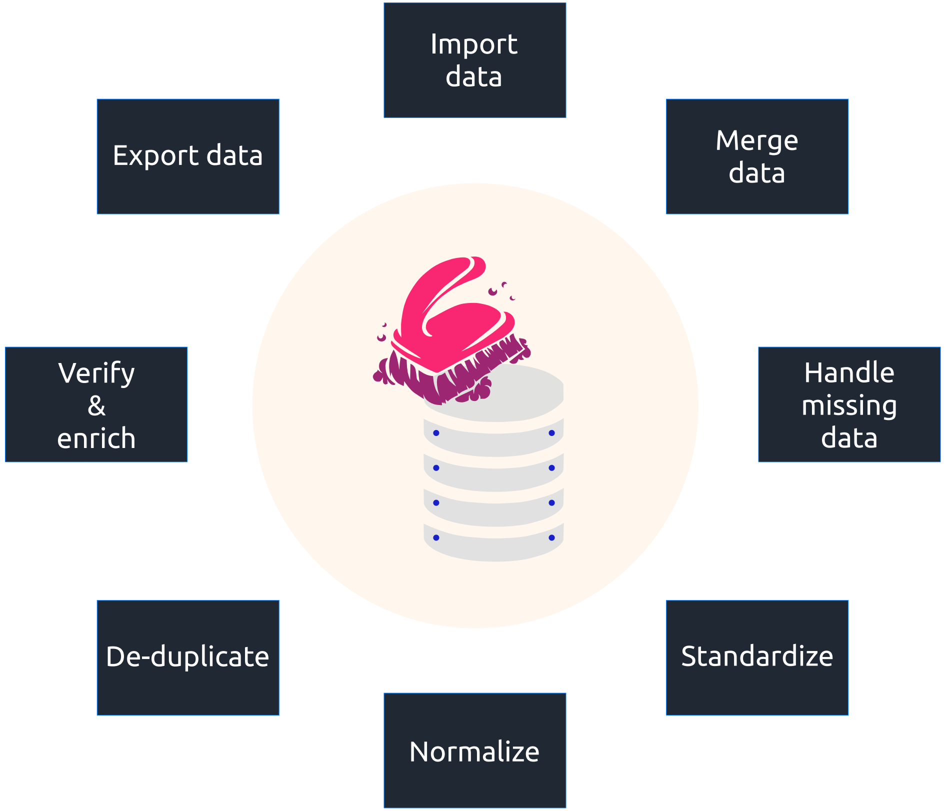


Use forms to improve quality

Name	Surname	Sex	Birthday	Birthplace	Country	Phone
Max	Rockatasnky	M	10-11-1984	Perth	AU	+61 8 6245 2100
Immortan	Joe	m	01-02-1949	Canberra	AU	+61 4 1234 5678
James	Connor	M	1985-02-28	Los Angeles	USA	unknown
Alex Murphy		M	1979	Detroit	US	tbc
John	McClane	M	1969-07-17	Los Angeles	US	4242706247
Pete	Mitchell	MALE	1972-10-10	San Diego	US	tbc

Processing

Cleansing


Import data
Merge data
Handle missing data
Standardize
Normalize
De-duplicate
Verify & enrich
Export data

But... what is exactly cleansing?

Activity	Example
Import	Retrieve data from DB, files, web scraping, APIs...
Merge	Combine data, combine tables by indexes, by column values...
Handle missing data	Remove entries, substitute with similar values...
Normalize	Numeric: Rescale values into [0, 1]
	NLP: Tokenization, Lemmatization, Sentencing...
Standardize	Rescale data to have $\mu = 0$ and $\sigma = 1$
De-duplicate	Drop duplicates
Verify and enrich	For dates, check dates follow the calendar and convert types
Export data	Save in a DB, in a file... (formatting)

Data

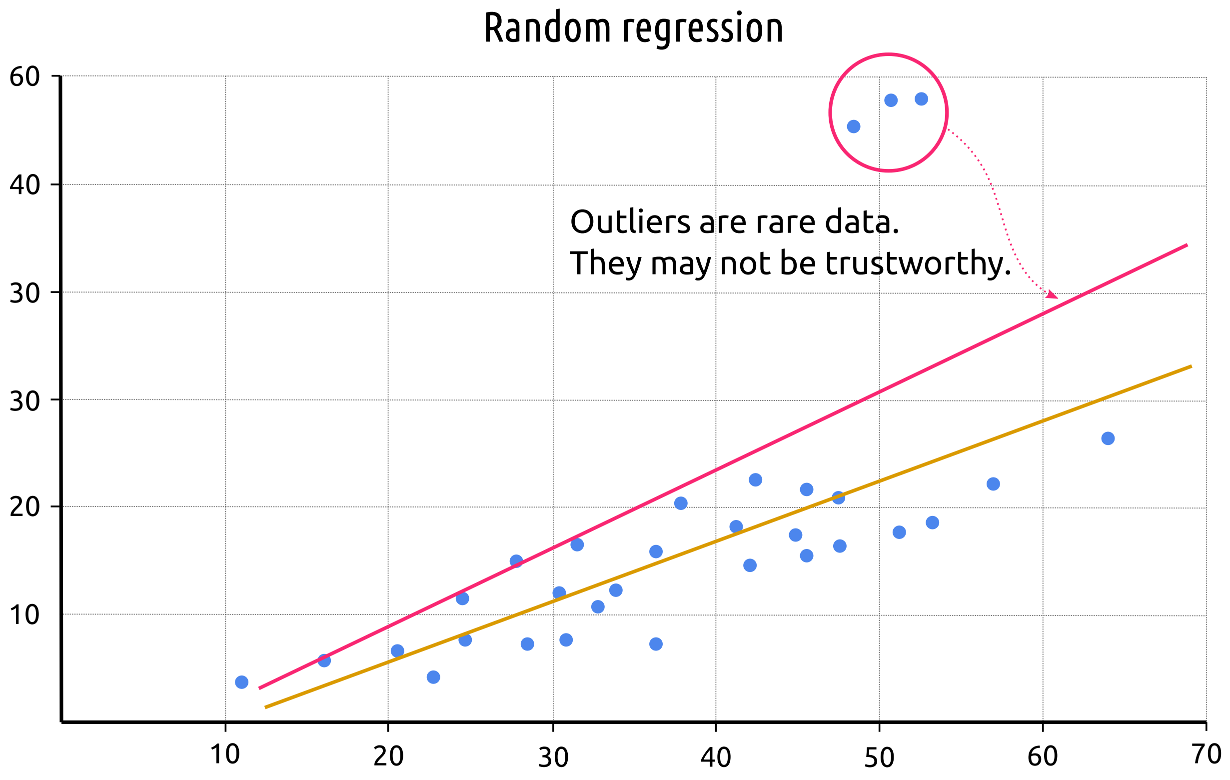
Processing

Feature extraction

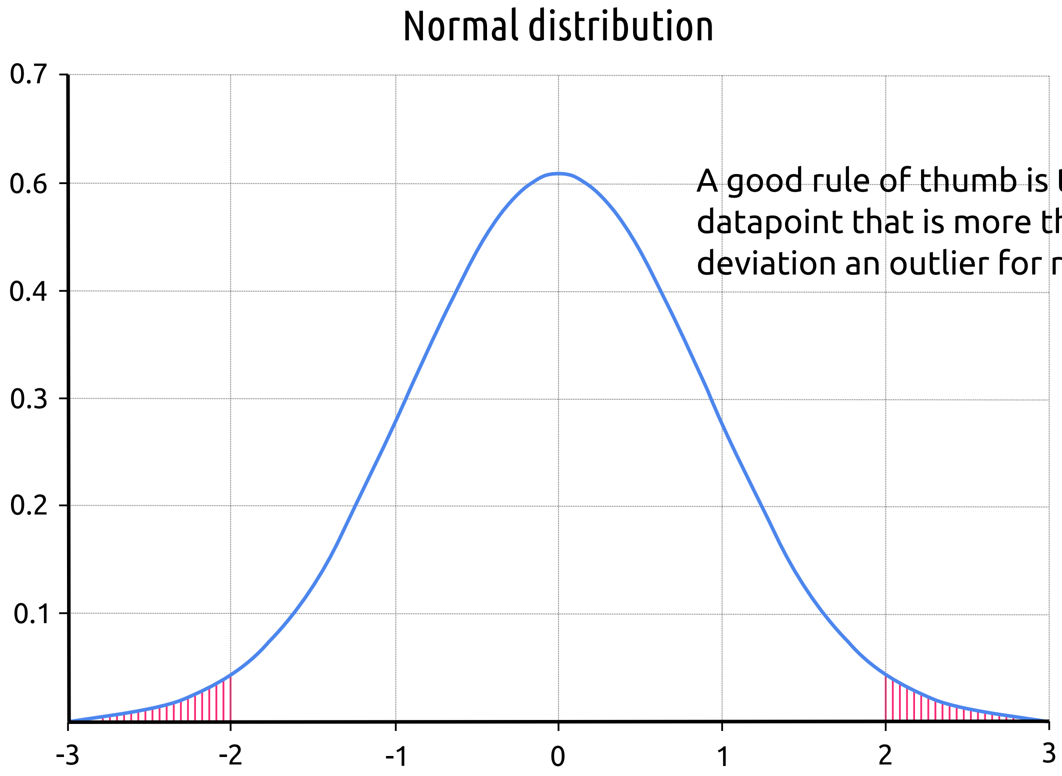
Modelling

Results

Consider removing non-representative data



Consider removing outliers



A good rule of thumb is to consider any datapoint that is more than 2 standard deviation an outlier for removal.

Feature extraction

The bridge

From Processing

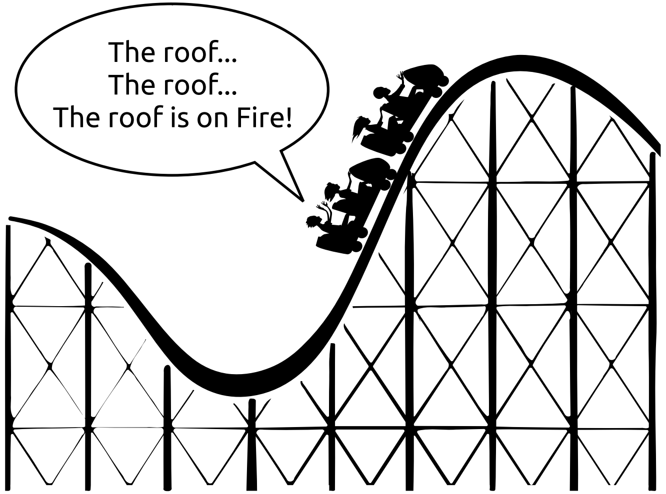
- Original data
- Clean and processed
- Not always informative
- Can be redundant
- Multi-dimensional

To Modelling

- Derived data· Inform

- Less dimensions

The "bridge" is not as it sounds...



Feature extraction

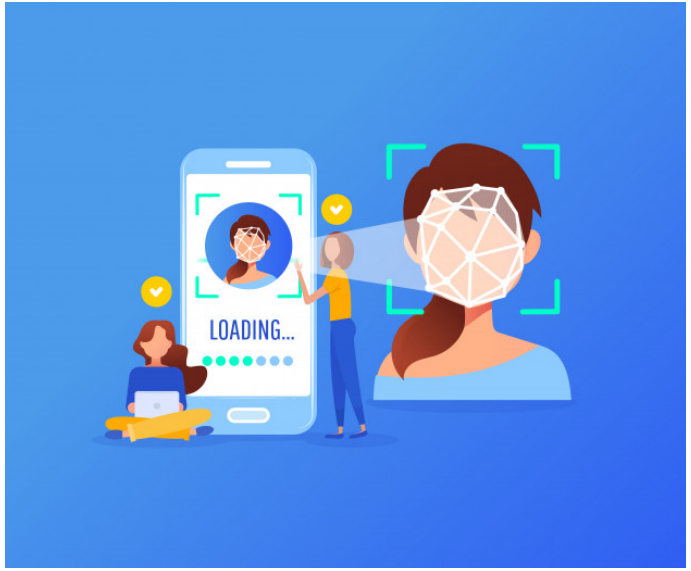
From Processing

· Original data
· Clean and processed
· Not always informative
· Can be redundant
· Multi-dimensional

To Modelling

· Derived data
· Informative· Less red

Example: facial recognition



Facial recognition applications extract key positions from your face and then:

- Calculates distances
- Calculates color

Your face becomes a vector of features!

Feature extraction applied to natural language

Natural language text cannot be used in any algorithm as it is. It must be converted to numbers:

There are several techniques involved

- Tokenization
- Lemmatization
- Stemming
- Vectorizers:
  - CountVectorizer
  - TfIdfVectorizer

All these techniques will be reviewed during the hands-on session.

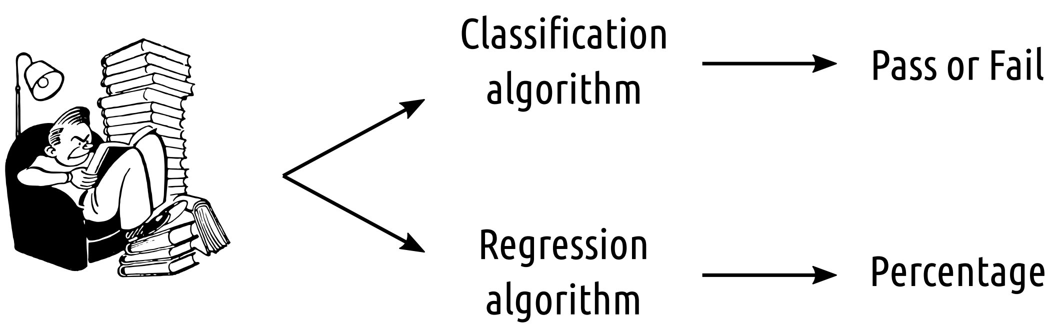
Modelling

Supervised learning / Unsupervised learning

Supervised learning	Unsupervised learning
Built on... Knowledge of desired output	**Built on...** Patterns or structures identified in data (no knowledge of output class or value)
Dataset Labelled (class or value)	**Dataset** Unlabelled
Goal Predict label (class or value) of data points	**Goal** Identify groups of similar data points based on internal criteria
Main applications Classification, regression	**Main applications** Clustering

Supervised learning

Example of supervised learning: the student



Features					Labels	
Name	**Class**	**Assistance**	**Course year**	**Hours/Week**	**Classification**	**Regression**
Mike	Maths	Yes	4	8	Pass	70%
Sara	Maths	Yes	4	12	Pass	99%
Paul	English	No	4	2	Fail	30%

Data

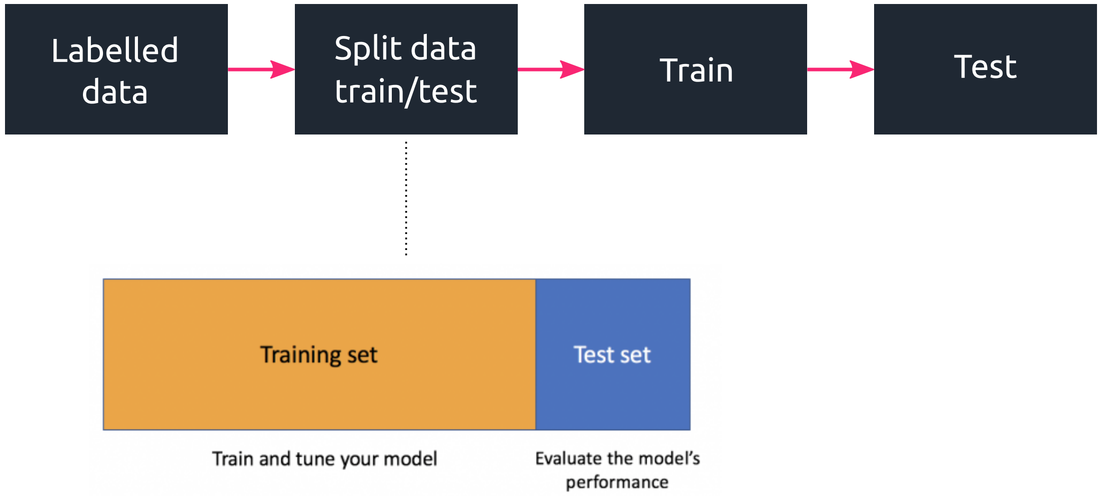
Processing

Feature extraction

Modelling

Results

Supervised learning pipeline



Data

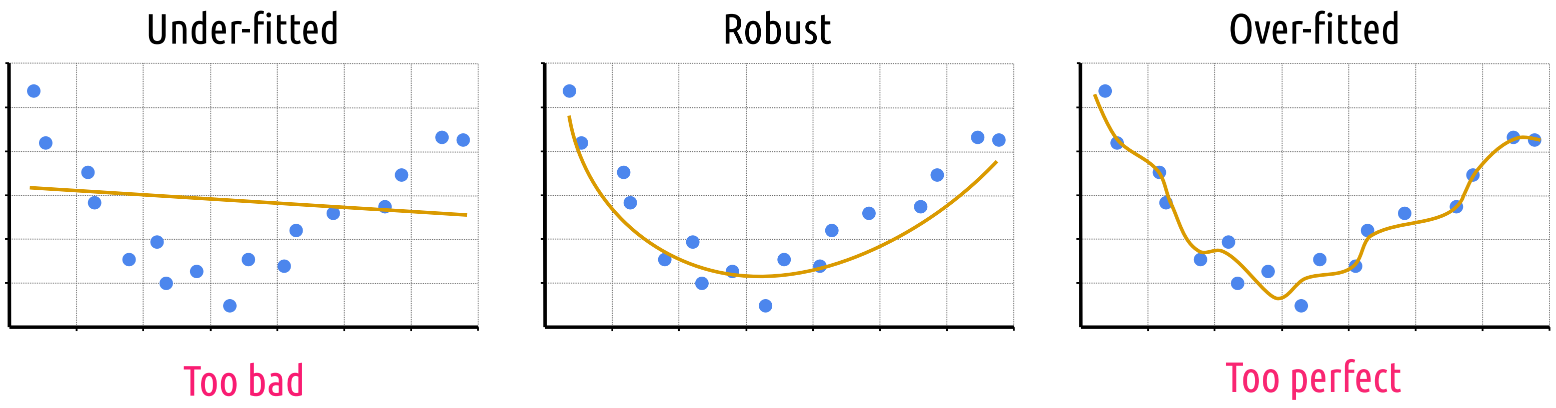
Processing

Feature extraction

Modelling

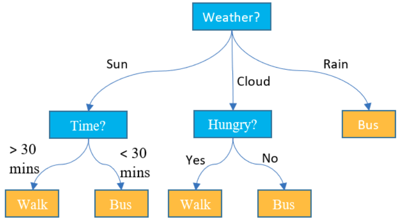
Results

Under-fitting / Robust / Over-fitting



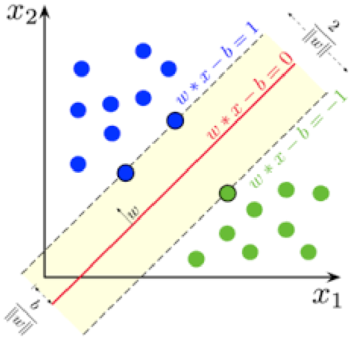
Some supervised learning algorithms

Decision Trees



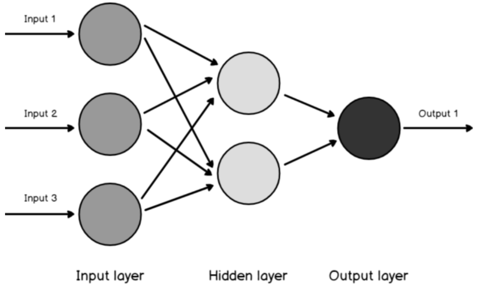
Easy

Support Vector Machines



Medium

Neural Networks



Difficult

Decision Trees

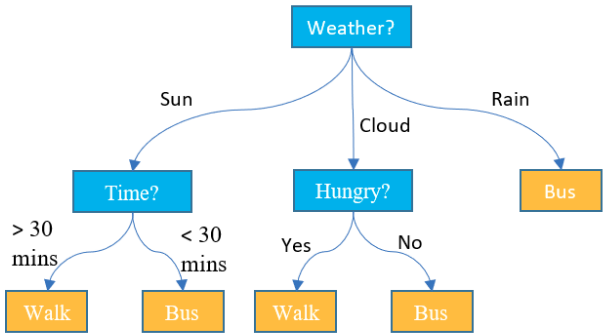
What

It is a versatile supervied algorithm for both classification and regression tasks.

How

It is based on trees:
· Each node is a condition
· Branches are the result of the conition.

Normaly trees are binary trees (i.e. two branches per node), but some adaptations (like ID3) can produce multiple branches.



Decision trees

· Require little data preparation
· Produces visual results easy to understand
· The depth of the tree is configurable

Data

Processing

Feature extraction

Modelling

Results

Support Vector Machines

What

Versatile algorithm that can perform linearand non-linear classification, regression tasks and outlier detection.

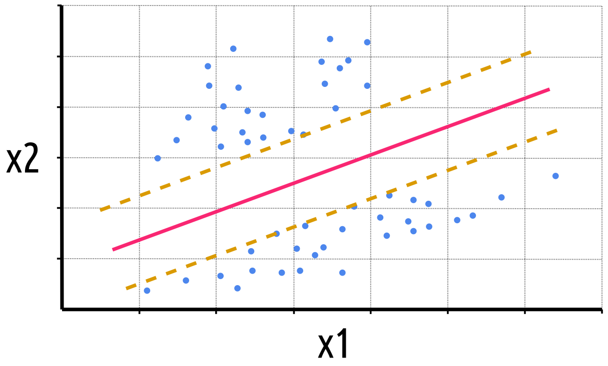
How

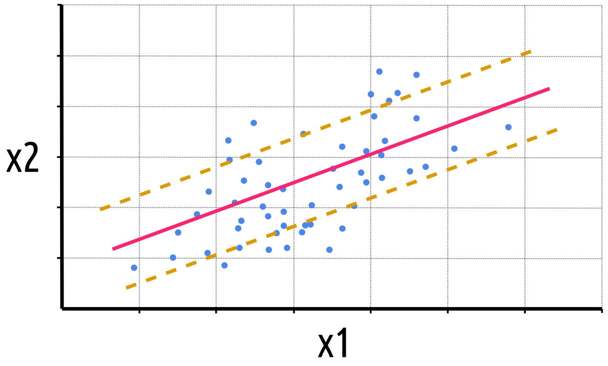
For classification: finds hyperplanes (i.e. streets) to separate groups of data.

For regression: finds "streets" whith as much data as possible in it.

Benefits

· Handles well unbalanced data
· Resistant to overfitting
· Works very well when identifying boundary regions





Data

Processing

Feature extraction

Modelling

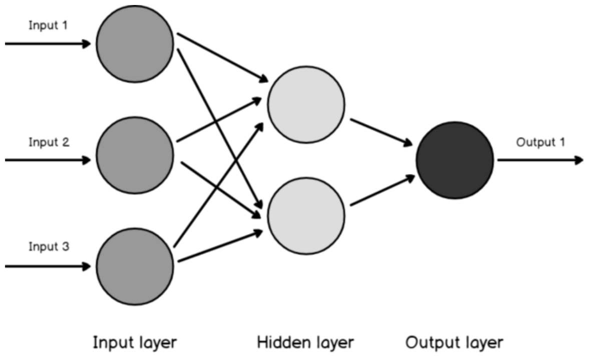
Results

Neural networks

What

It is a classifier modeled loosely after the human brain designed to recognize patterns.

It has a set of neurons (perceptrions) organized in the form of multiple hidden layers, lying between the input layer (input data) and the output layer.



How

Networks are very useful in scenarios where the relationship between input features and output classes appears vague.

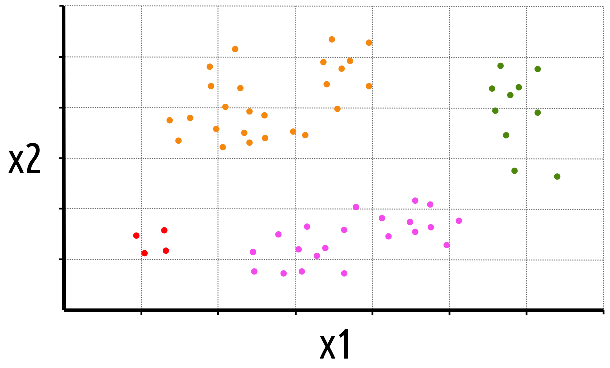
Benefits

· Black box
· User only has to configure the NN (layers, perceptron/sigmoid...)
· Ideal for high-dimensional datasets

ZURICH | Technology Delivery Center ServiZurich I Barcelona

Unsupervised learning

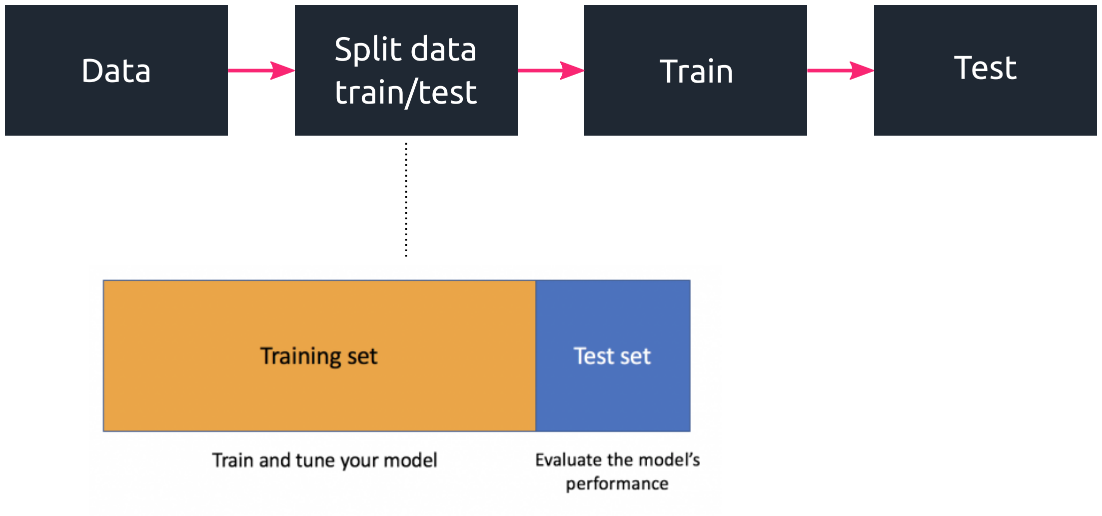
Example of supervised learning: clustering video-games clients



Features

Name	Age	Sex	Num. kids	Income	Weight	Smokes
Mike	18	M	0	18000	74	1
Sara	29	F	0	56000	56	1
Paul	43	M	1	49000	82	1

Unsupervised learning pipeline

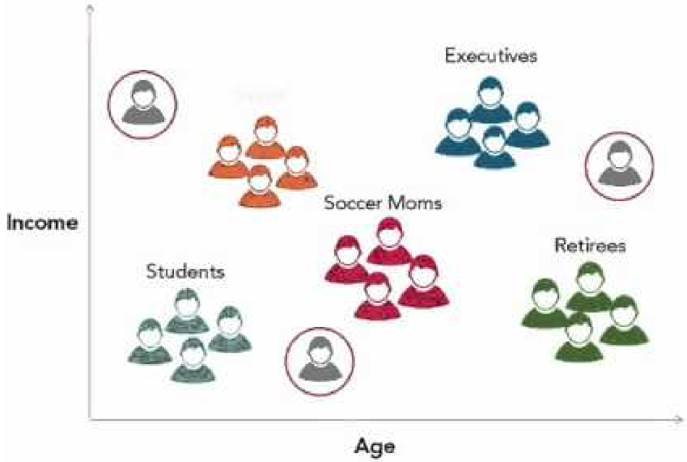


Clustering

What

Clustering is one of the most common forms of unsupervised learning (k-means and hierarchical clustering).

Clustering, is used primarily to:
· Segment data
· Detect anomalies
· Simplify datasets by aggregating variables



How (k-means)

Group by similarity.

How (hierarchichal)

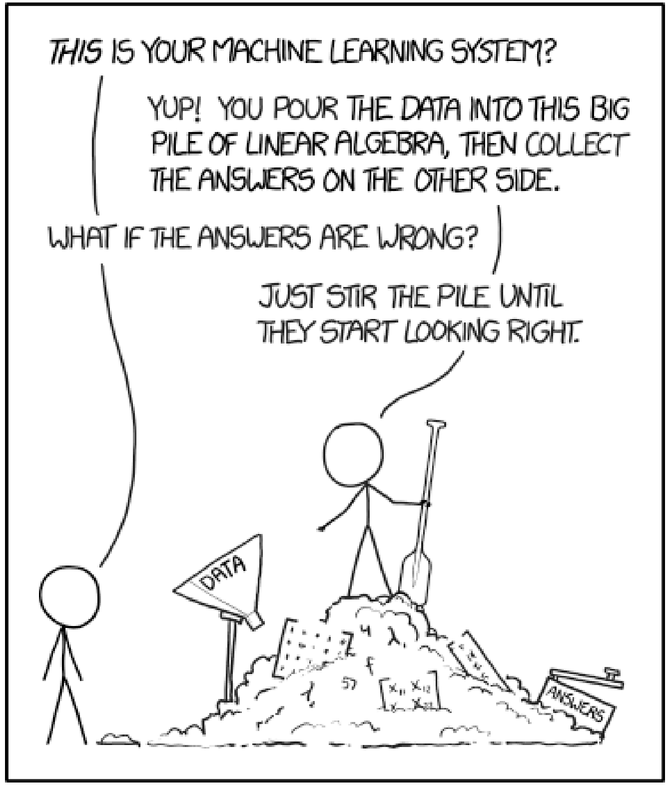
Group all data in one cluster. Then split it until all of the have one sample

Benefits

· Handles well unbalanced data
· Resistant to overfitting
· Works very well when identifying boundary regions

Results

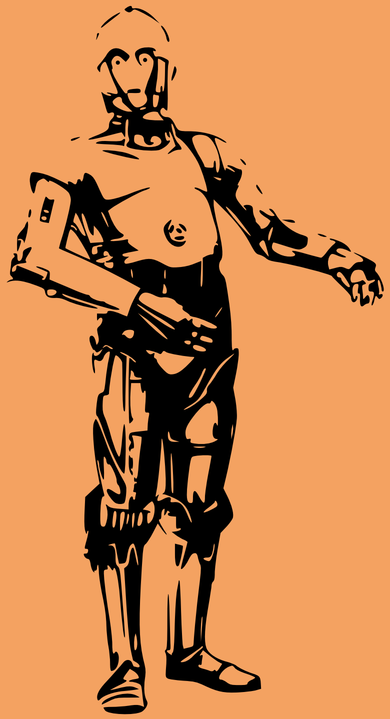
Results


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Hands-on ML (practice)

Go to:

https://colab.research.google.com



Questions?

(albert.ruizalvarez@zurich.com)

Thank you!